Name: Nishant Nitin Musmade

Date: 13/09/2024

Title: Vehicle Data Analysis

Table of Contents:

Introduction:

The goal of this task is to conduct exploratory data analysis on the provided vehicles dataset, and to gain some valuable insights into various aspects of vehicles attributes and trends. The analysis will further help to reach the potential areas which can be studied and improved. Key steps include data preparation, cleaning, EDA and feature engineering.

Data loading and Preparation

Dataset was initially loaded, and its structure was inspected, where the shape of the dataset was (18434 x 141). On checking dataset information, we come to know that initially it had 122 float64 datatype, 13 int64 datatypes, and 6 objects. After checking for the null values in the dataset, some columns are found to be completely empty. So, to make further process simple, these columns were deleted as it will not contribute to any analysis process. In addition to this, the dataset contained columns that has same value for all records, so these columns were also deleted.

Feature 'can_raw_data' and 'pluscode' are also dropped from the dataset because 'can_raw_data' is not interpretable as we don't have proper documentation to decode the CAN message and in case of pluscode which is used to extract latitude and longitude, the information that we already have in other columns named as 'lat' and 'lng'.

Data cleaning and preprocessing:

There were almost 0% missing values in each column except 'adblue_level', so dropna() method is used to drop the records containing missing values. For 'adblue_level', approximately 19% missing values were there. For handling these missing values, a machine learning model named 'RandomForestRegressor' was used. For this model, a training set was created which was a subset of the original dataset with no missing values. Then the model is trained on this subset and it achieved a good R score of 99.83% and finally this model was used to predict the missing values in the original dataset.

Also, the dataset consists of three categorical variables (clutch_switch_status, brake_switch_status, parking_switch_status). Label encoding was applied on these three features (0-releases, 1-pressed). The dataset also consisted of outliers so the skewness of each

feature was examined and data transformation was done wherever necessary using the methods such as log transformation(right-skewed) and box-cox transformation(left-skewed). Then we plotted a box plot to identify outlier as shown in fig.1 with the help of Interquartile Range (IQR) method and then the outliers were removed from the features.

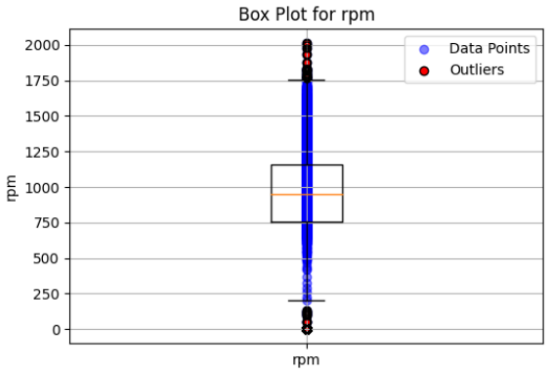


Fig.1. Box plot for rpm

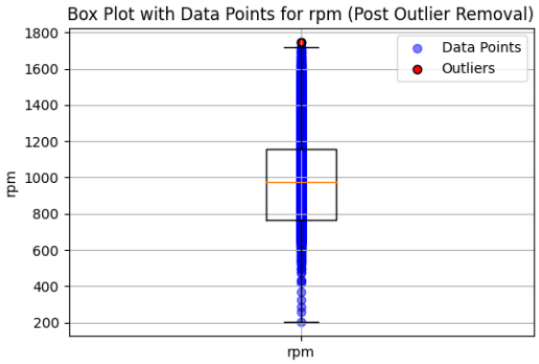
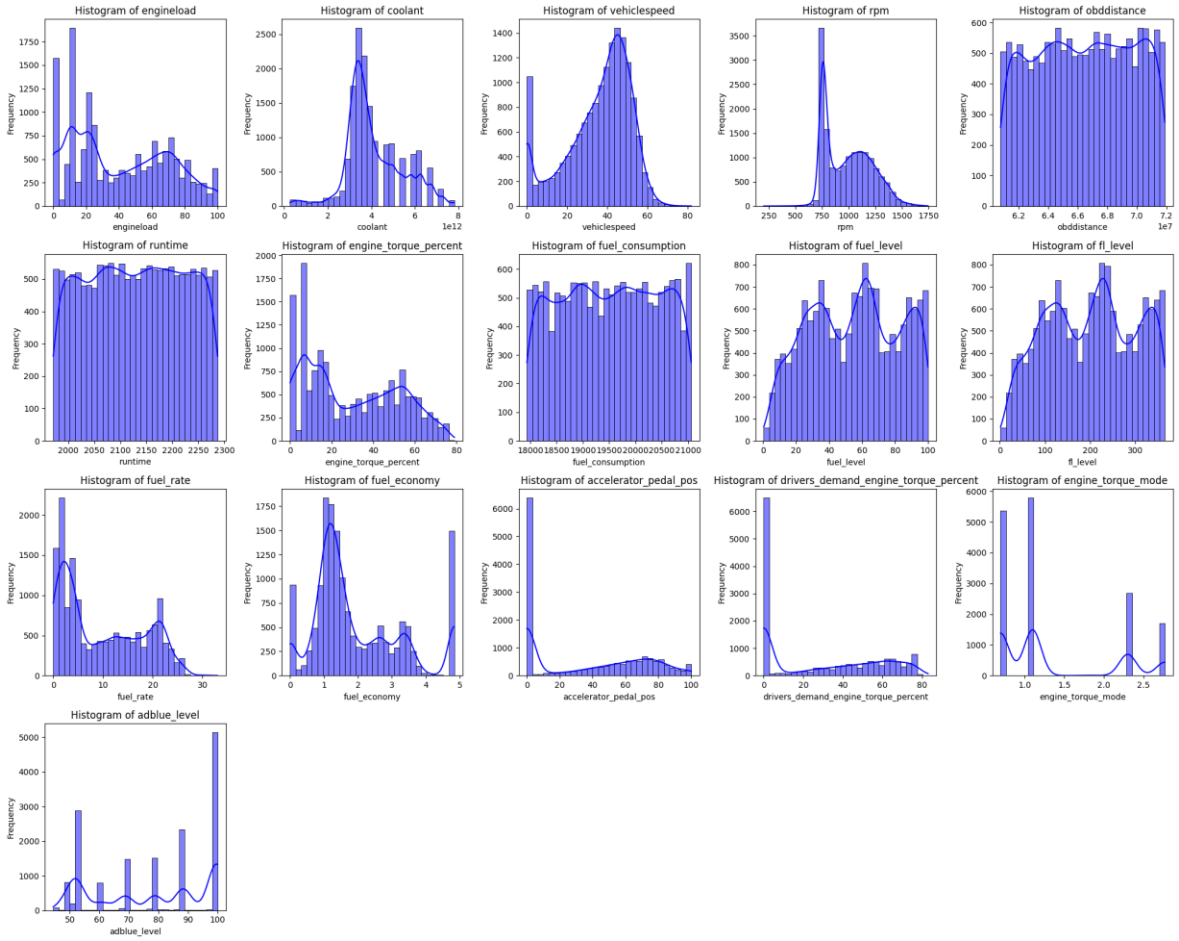


Fig.2 Box plot after removing outlier

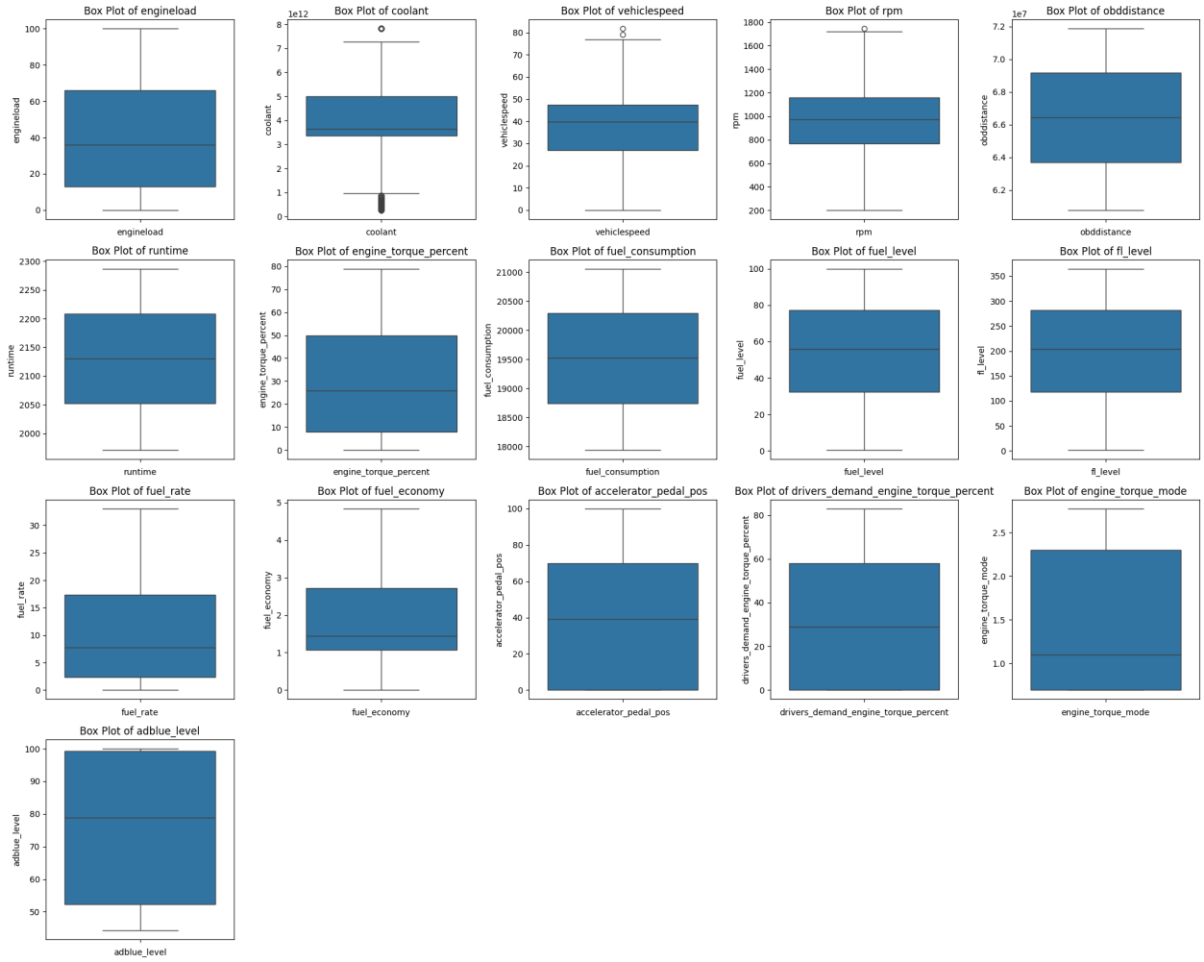
After the data cleaning process, the shape of the dataset was (15586 x 24), and now it has no missing values and very less outliers. The data is now ready for the EDA process

Exploratory Data Analysis (EDA):

1. Histogram for visualizing the distribution of data



2. **Box plot to identify the outlier, and visualize the spread of data, including quartiles and median.**



3. **Scatter plots to visualize a relationship between two numerical variables, helping in identifying trends, correlations, and outliers**

   a. **engineload vs fuel_rate:**
   - The scatter plot suggests a positive correlation between engineload and fuel_rate. As the engine load increases, the fuel rate also tends to increase.
   - This plot can be useful for understanding how the engine load affects fuel consumption. Higher engine loads seem to correspond with higher fuel rates.

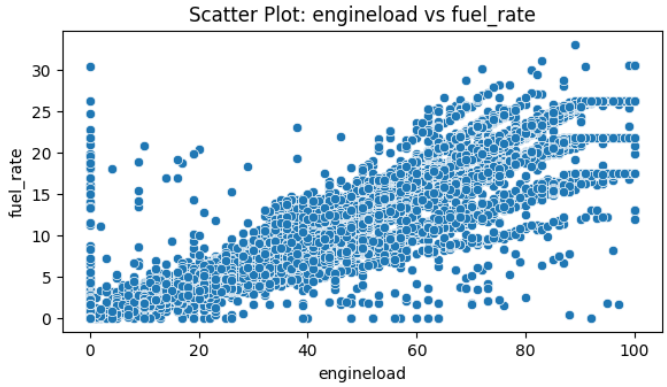
   b. **Vehiclespeed vs rpm:**
   - The scatter plot suggests a positive correlation between vehiclespeed and rpm. As the vehicle speed increases, the rpm also tends to increase.
   - This plot can be useful for understanding how the vehicle speed affects engine performance. Higher vehicle speeds seem to correspond with higher rpm.

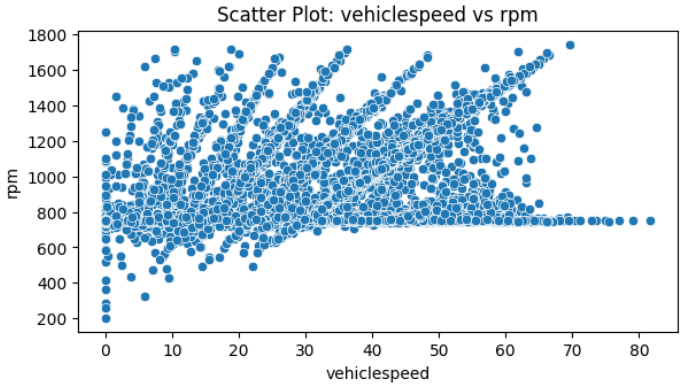
c. engine_torque_percent vs fuel_rate:

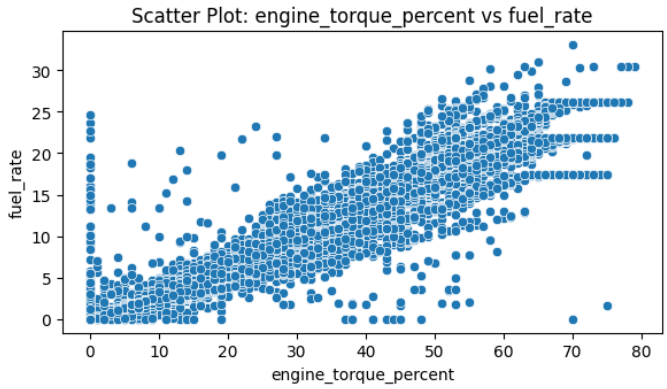
- It is also indicating a positive correlation between engine_torque_percent and fuel_rate, that means higher torque percent tends to increases fuel rate.

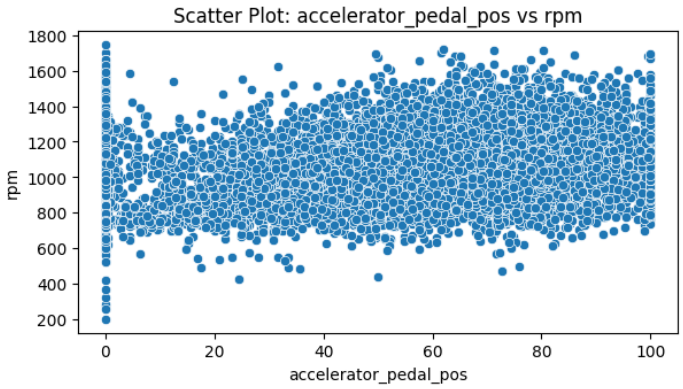
d. Accelerator_pedal_pos vs fuel_rate:

It is also indicating a positive correlation between accelerator_pedal_pos and fuel_rate. This suggests that accelerator_pedal_pos also affects engine performance.



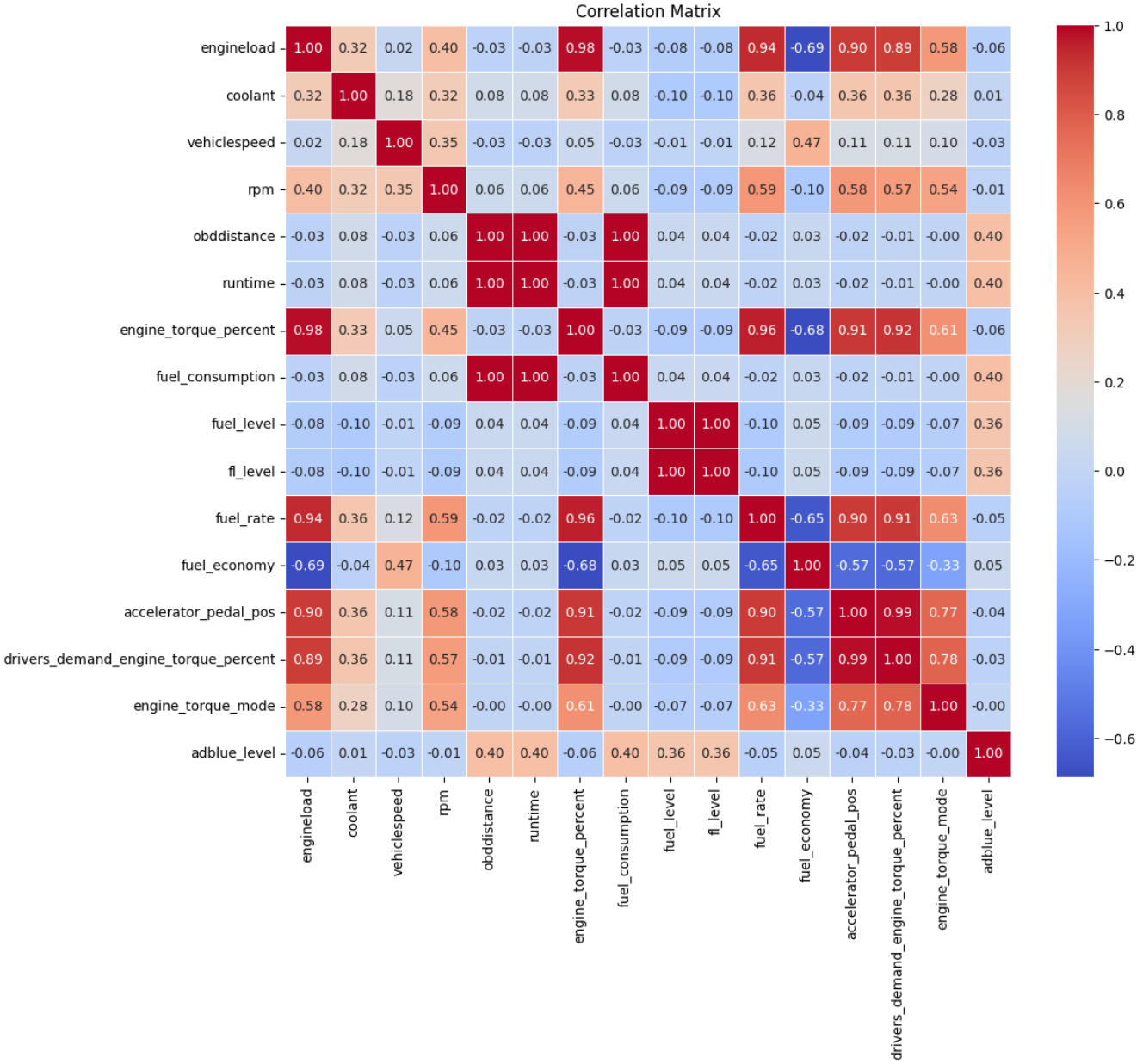






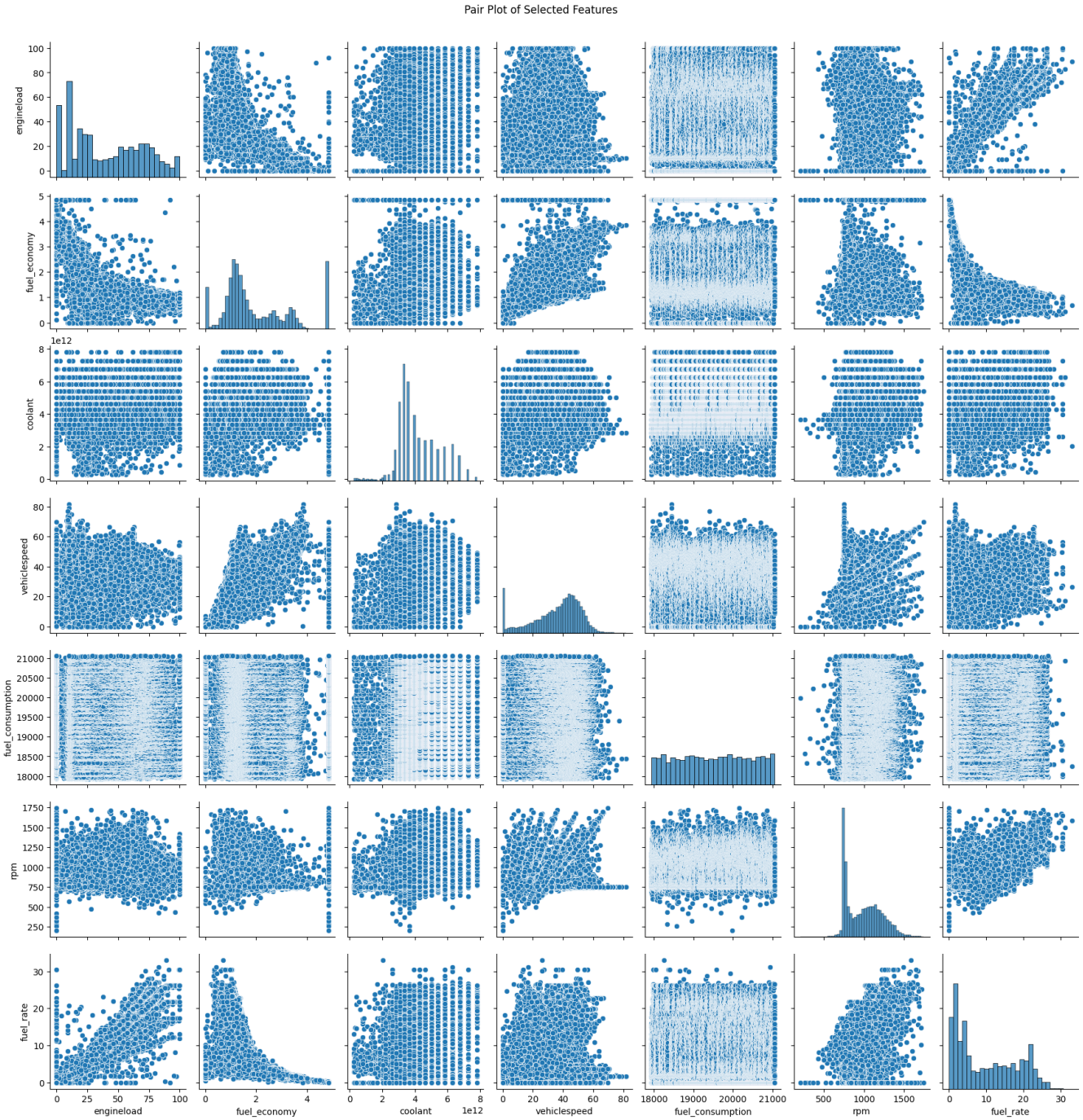
4. **Correlation matrix:**

The correlation matrix shows the pairwise correlation coefficients between numerical variables. This helps to understand the strength of relationships between variables.

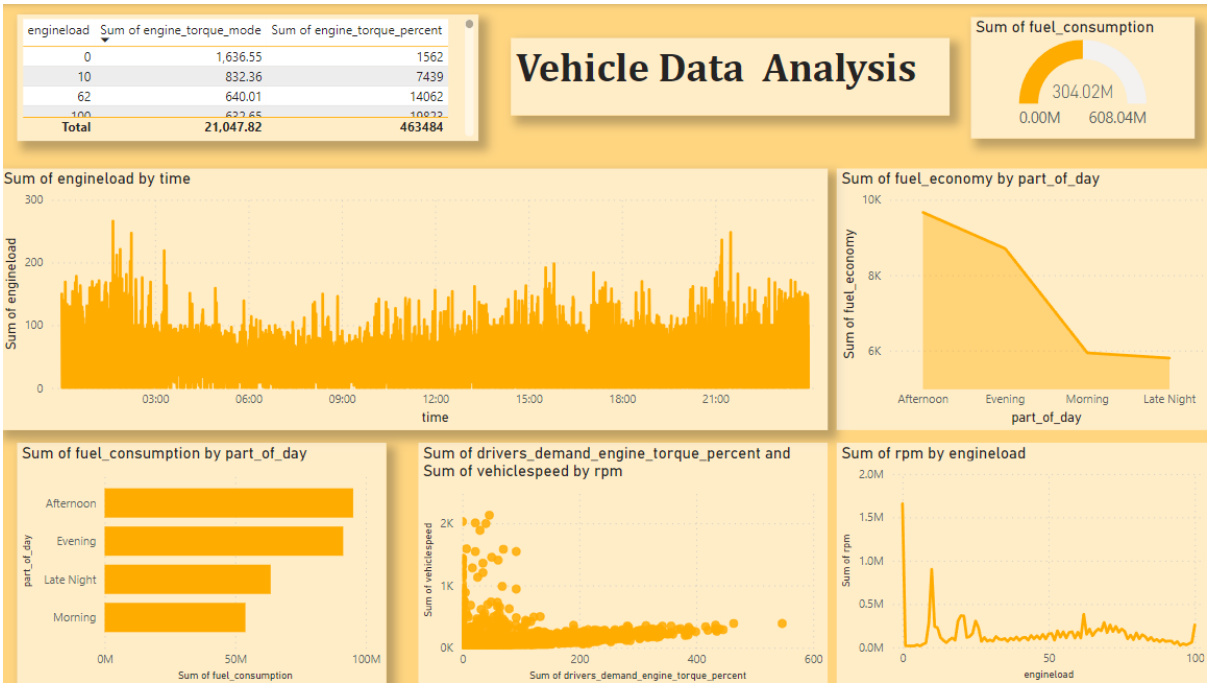


5. **Pair plot:**

Pair plots show scatter plots for each pair of features, as well as histograms for individual features. This helps to visualize relationships and distributions for all pairs of features. Here, only important feature that are suitable for pair plot is selected as it become difficult to interpret if all features were included.



6. **Dashboard using PowerBI:**



Feature Engineering

a. **Creating new features from existing data**
One of the features was timestamp 0(ts) which is not much readable, so from 'ts' two features were extracted – date and time, which is more readable as compared timestamp. In addition to this, we have extracted different parts of the day as Late night, Morning, Afternoon and Evening. This could help for analysing trends in vehicles on the basis of different parts of the day.

b. **Normalizing features:**
Features such as rpm, obddistance, runtime,fuel_consumption, and fuel_rate are normalized as these features have range of values which are different as compared to other features. For normalizing MinMaxScaler() is used which scales the data to fixed range (0 to 1).

Insights and recommendations

a. Insights:

1. Fuel Economy Patterns: Vehicles with higher engine loads tend to have lower fuel economy, indicating the relationship between engine load and fuel efficiency.

2. Coolant and RPM Relationship: Higher RPMs increase coolant temperatures, showing a strong connection between engine speed and thermal performance.

3. Engine Load and Fuel Rate: As engine load increases, fuel rate increases proportionally, indicating efficient fuel delivery under load conditions.

4. Skewed Distributions: Several features, such as fuel economy, coolant, and engine torque mode, were highly skewed, requiring transformations to normalize the data for better analysis.

5. Outliers Detected: Extreme values in features such as RPM and vehicle speed were detected as outliers and removed to ensure more accurate modeling.

6. Fuel consumption in a day: The afternoon is the time when fuel consumption is highest and, in the morning, fuel consumption is lowest.

b. Recommendations:

1. Fuel Efficiency Optimization: To improve fuel economy, it is recommended to monitor and limit engine load by optimizing driving habits.

2. Engine Performance Monitoring: Regular monitoring of coolant temperatures and RPMs can help detect engine overheating or excessive wear.

3. Driver Behavior: Encourage drivers to maintain optimal speeds and throttle positions for fuel efficiency, particularly on long trips.

4. Outlier Detection in Real-Time: Areal time monitoring system can be implemented to detect anomalies in RPM, engine load, and speed, which can indicate potential vehicle issues at the earliest.